Issues on East Asian Character Codes and Unicode: What Happened to the Character I Input?

Kuang-tien (K.T.) Yao University of Hawaii at Manoa

Introduction

After more than two years of planning on the Endeavor Implementation Project, in January 2001, University of Hawaii at Manoa (UHM) Library and thirteen other libraries in the State of Hawaii finally launched a new client-server based Voyager System designed by Endeavor Information Systems, Inc. (EISI). Several months before launching Hawaii Voyager System (version 99), EISI announced that its next product enhancement would be based on "glyph server" technology, developed jointly with InterPro Global Partners. The glyph server takes bibliographic information stored in the MARC standard format (MARC-8), converts it to Unicode standard characters in UTF-8, and then publishes the characters in the form of a language-specific set of glyphs—images that are viewable from any standard Web browser. The Glyph server allows the OPAC end-user to accurately see the language represented without having to download font sets for each language.[1]

EISI's development for the Unicode language features have created an opportunity for UHM to be involved in an cooperative development to work together to design and implement support for Chinese/Japanese/Korean (CJK) and other non-Roman character sets within the software. The focus of these enhancements is the support of these non-Roman character sets within bibliographic records through the use of Unicode. For the development of Unicode™ capabilities in the 2000.1 release, EISI has formed an Unicode™ Task Force which brought together ten representatives,[2] representing diverse users from various libraries, to work with its software development teams on the development of database conversion, public display of non-Roman scripts, input of non-Roman text in Cataloging module, and record import and export. Among the ten representatives who have participated in this project, seven are network and system experts and three are Chinese librarians – Martin Heijdra,[3] Zhijia Shen[4] and I. Since my primary responsibility at the University of Hawaii Library is cataloging Chinese materials in all formats, my participation on the Unicode™ Task Force focuses mostly on CJK display on OPAC.

In this article, I would like to talk about my experience on some issues and problems, particularly on EACC and Unicode that I have encountered while working on the

---

[1] "Endeavor offers Unicode capabilities." *Library Automation.* http://www.infotoday.com/cilmag/oct00/newsline.htm

[2] The ten representatives are from Cambridge University, Cornell University, Getty Research Institute, the University of Hawaii, the Library of Congress, the Linnea2 Consortium of Finland, the University of Pittsburgh, Pepperdine University, Princeton University and Yale University.

[3] Martin Heijdra is Chinese Bibliographer and Head of Public Services for the East Asian Library of Princeton University.

[4] Zhijia Shen was the Head of East Asian Library at the University of Pittsburgh. Currently she is the Head of East Asian Library at University of Colorado at Boulder.

Unicode™ Task Force for reviewing CJK display. I hope this article will alert my colleagues, particularly catalogers who input CJK records on RLIN or OCLC databases that build on the East Asian Character Code (EACC), to be aware of the complexities of selecting Chinese characters while creating or inputting bibliographic records.

University of Hawaii Library at Manoa is a RLIN contributor. CJK cataloging staff members contributes our CJK bibliographic records to RLIN national database via the internet. As one of the catalogers at the library who uses RLIN® Terminal for Windows to input Chinese MARC records daily, I am confident of my understanding and familiar with this national utility. However, after working on the Task Force, I realize that even though I may be familiar with using the RLIN Terminal for searching and inputting bibliographic records, I am less aware of certain complications in the EACC of CJK characters in the RLIN thesaurus, not to mention a basic understanding of the Unicode Standard. In order to better review the CJK display, I began reading some basic books about Unicode and surfing web sites to find more information about EACC and Unicode. One very useful source is EACC/Unicode Review Project,[5] a CEAL project led by Bob Fleshing, based on Unicode version 2.0, which has helped me to understand some of the issues and problems I have encountered.

With my cataloging experience, language skills, and a basic understanding of EACC and Unicode, I worked with Zhijia Shen, Martin Heijdra and other system experts to review and debug the CJK display on the Voyager's preview glyph server, and we were able to point out problems of Endeavor's software development teams. Consequently, some CJK display problems were resolved and some EACC/Unicode mapping codes were revised. However, there were still unresolved issues on the characters display. For example when an EACC character, which lacks Unicode, is mapped to a Private Use Area (PUA) code point, it will cause the disappearing of that character on OPAC because PUA (ranged from E000-F8FF) doesn't contain any character assignments.

Missing Characters: Why Don't They Display?

Missing characters were the biggest surprise to me when I reviewed the CJK display. I asked myself, "Didn't I just finish a complete record on RLIN with all the Chinese characters? I even have my printout to prove that I did, but why did they not display on OPAC?" To solve this mystery, I checked several sources, including a RLIN thesaurus, web page sources for that problem record,[6] the Unicode Standard Version 3.0,[7] and finally, EACC/Unicode CJK mapping files on our Voyager system.[8] From the RLIN thesaurus I was able to find the EACC. The Unicode Standard provided me with the Unicode and its matching ideograph. The page source shows the Unicode used for the

[5] http://fluffy.uoregon.edu/unicode/

[6] Page source for the bibliographic record can be viewed from any web browser. Just go to [View] then select [Page source].

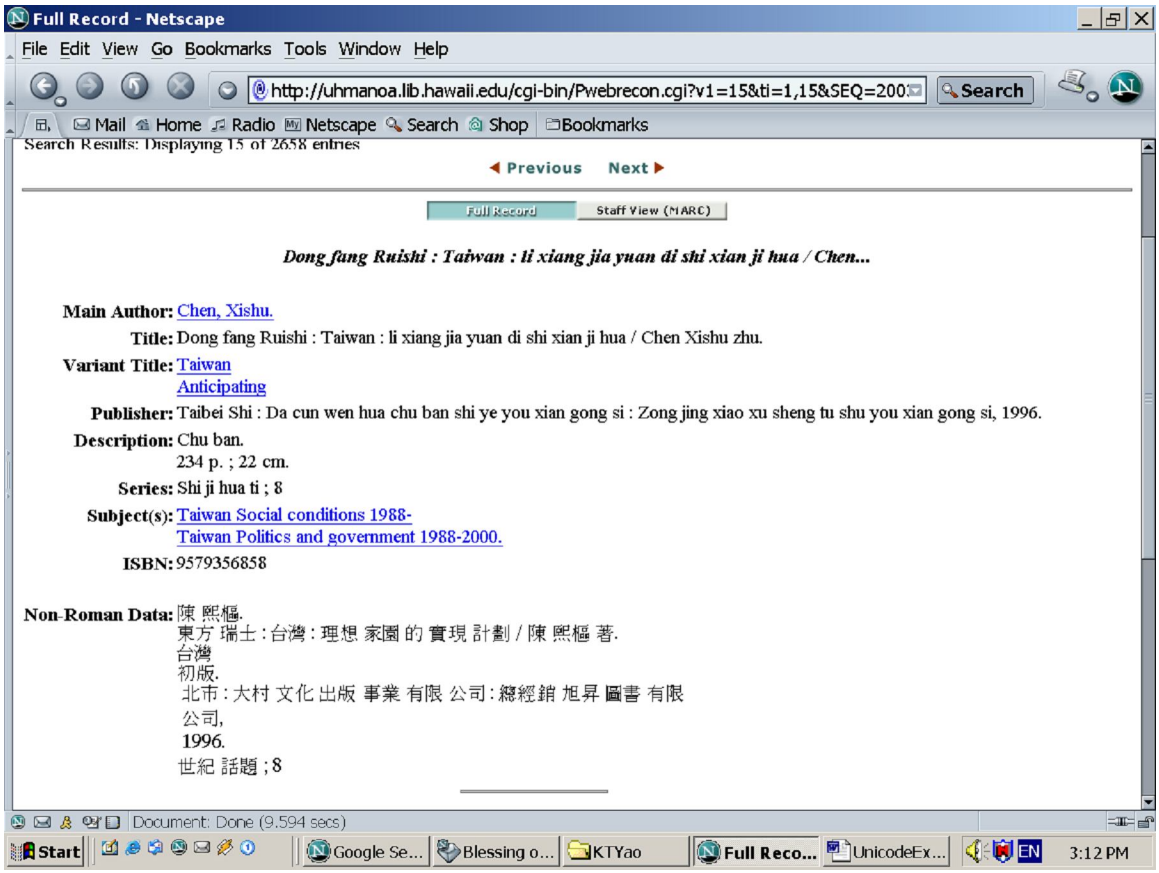
[7] *The Unicode Standard, version 3.0*. Reading, Mass: Addison-Wesley, 2000.

[8] Voyager system's CJK mapping file is based on the EACC/Unicode mapping approved by MARBI Committee of the American Library Association, which is also available on the MARC21 website http://www.loc.gov/marc/specifications/specchareacc.html

characters. After finding the EACC and carefully identifying the Unicode, I checked Voyager's EACC/Unicode mapping file to make sure that there is the EACC and that it matches the appropriate Unicode. Finally, the mystery was revealed.

Why Some Display But Some Don't?

The main reason for the disappearing of certain characters is because not all CJK characters represented in EACC have Unicode values, even though the Unicode Consortium claims "Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language." [9] Here is a typical example showing that one simplified Chinese version of "tai (台)" (in 台北市) is missing, but two others (台 in 台灣) in the title and variant title display.



For this particular example, it is because the cataloger used two different "tai" which looks identical on RLIN but coded in two entirely different EACC values in the same record in RLIN. After importing this RLIN record to Voyager system, the display one (coded EACC: 27542B) has mapped to an appropriate Unicode (53F0) in the Voyager CJK mapping file. The non-display one (coded EACC: 213538) has mapped to the PUA

[9] Unicode Consortium website: http://www.unicode.org/standard/WhatIsUnicode.html

(E900). Since PUA doesn't contain any character assignments, the character doesn't display on the OPAC.

I Thought I Had Selected the Correct Character for My Record on RLIN

There are approximately 297 variant character forms that have EACC values but actually lack Unicode. As a RLIN user, you are aware that when you type "tai" in the language input mode to retrieve a character (in this case is 台) needed for your record, the RLIN terminal provides the following three simplified versions of "台" for you to select.

Tai (台臺) (台台) (台颱)

Here is a simple table to give you the EACC/Unicode mapping information. Note: all three characters appear identical as 台 in RLIN Thesaurus

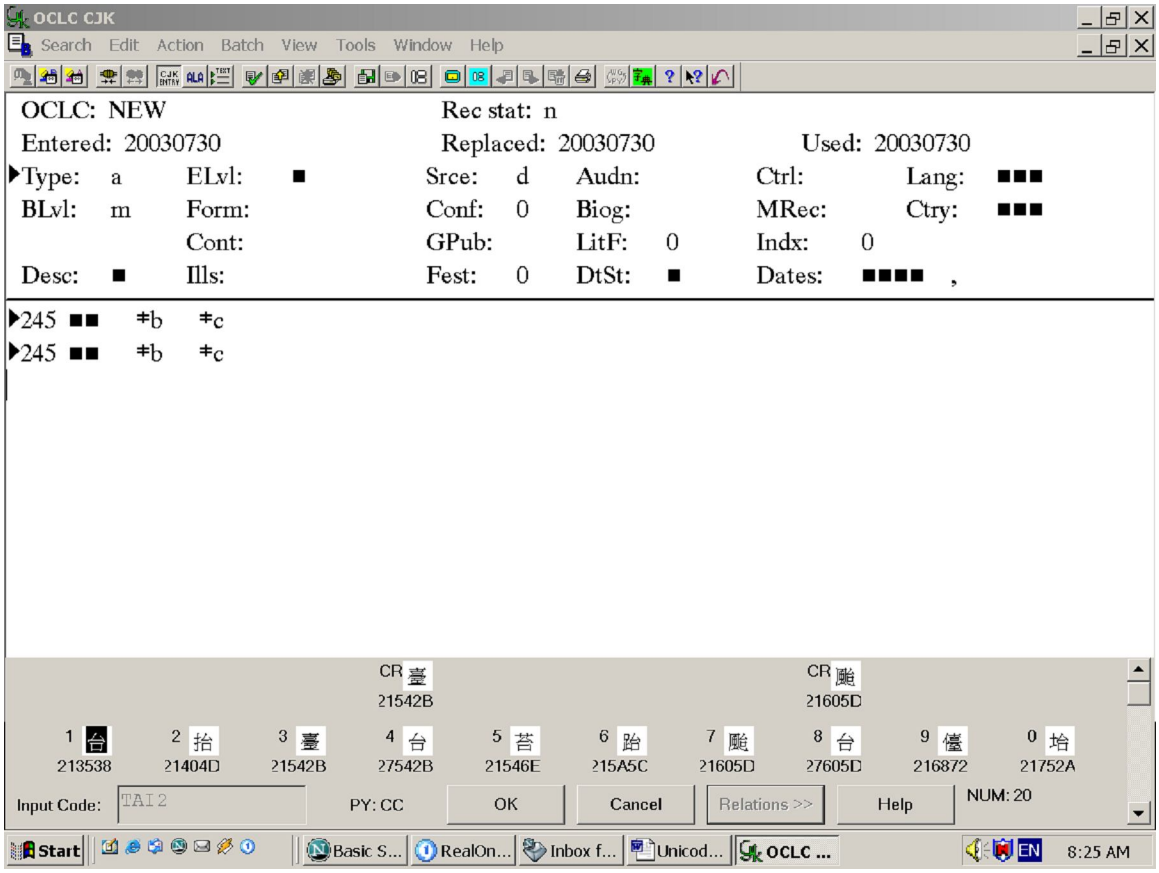
Character	EACC (Simplified Chinese) ; EACC (Traditional Chinese)	台 (Mapped to) Unicode
台 (台臺)	台: 27 54 2B ; 臺: 21 54 2B	53F0
台 (台台)	台: 21 35 38 ; 台: 21 35 38	E900 (PUA – doesn't display)
台 (台颱)	台: 27 60 5D ; 颱: 21 60 5D	E916 (PUA – doesn't display)

As you can see, "台" in the first group (台臺) is the simplified version of "臺." In the third group (台颱) is the simplified version of "颱." Amongst these three groups, only the first "台" in the first group (台臺) displays because it has mapped to an appropriate Unicode. If for some reason you have selected the second and the third "台" for your record, you have unknowingly created a problem, which is the missing character, on your record, because they are mapped to PUA codes in Voyager CJK mapping file.

For more detailed information on these characters, RLIN users may search RLIN CJK Thesaurus. OCLC users may check "Help Screen" on the OCLC CJK Passport or the website tutorial for *Learning to Use OCLC CJK Software (3rd ed.)*.[10] The website version has 14 lessons for you to learn the software. Lesson 2: *Creating CJK Characters: Phonetic Input Codes* provides a basic explanation on how to distinguish EACC values for simplified and traditional characters as well as useful information on how to work with identical characters. Also, it is very easy for OCLC CJK users to check "Relations" for variant characters on the CJK entry mode. All you need to do is to click on the "Relations" button when you retrieve some identical characters. It may be useful to check "Relations" to get some idea, at least, about these variant character forms.

[10] *Learning to Use OCLC CJK Software*: http://www.oclc.org/oclc/cjk/lessons/

Example of Identical Characters and Their Relations on OCLC CJK Entry Mode:



One other step a cataloger could do is to check the record display on your OPAC, especially for those identical characters that have different EACC values. Of course, it would be great if you could identify which EACC to use to avoid having missing characters, but this really requires familiarity with EACC/Unicode mapping.

Some Help is on the Way – Library of Congress' Alternate MARC8 to Unicode Mapping Characters

The good news is that the Library of Congress (LC) has begun working closely with EISI on its database conversion process. The first conversion of the LC database was done in January 2003. 31.7 million records, including 500,000 records with 880 tags, were converted to Unicode.[11] During the conversion process, LC and EISI learned that using the PUA mapping had affected over 10% of the EACC records, and that the overwhelming majority of PUA usage came from the variant character mappings with simplified Han, simplified Chinese characters made in Mainland China. As the result of

[11] *Voyager EndUser Conference Note*
http://staff.tuglibraries.on.ca/trellis/EndUser2003/EndUserGenUnicode.pdf

database testing, LC has put together alternative Unicode mappings for MARC21 characters assigned to the Private Used Area (PUA), which will be used in the Unicode conversion mapping. This new list of variant East Asian ideographs called *Alternate MARC8 to Unicode Mapping Characters*, which posted in June 16, 2003, is now available on the website: http://www.loc.gov/marc/specifications/puaalts.html. This document contains mapping for 297 MARC-8 EACC values that were mapped to the corresponding Unicode character values in the PUA and their alternative non-PUA character. They are divided into eight groups, including 151 variant Chinese ideographic characters, 28 Chinese characters not yet in Unicode, 38 characters that represent duplicate simplified Chinese ideographs, 5 Chinese ideographs not in the Unified Han set, 8 unrelated Chinese variant ideographs, 4 "Version J" Chinese ideographs (these are from 10 Chinese characters added to EACC prior to the implementation of UCS/Unicode), 28 ancient Korean hangul characters, and 35 component characters used in RLIN's proprietary CJK input method. LC's website is a very useful source for us to find out issues relating to PUA characters and how variant East Asian ideographs are mapped to their alternates.

Here is one of the groups that contains those 38 characters, including example for 台, that represent duplicate simplified Chinese ideographs. (Please visit the website for the complete list of all characters.)

Duplicate simplified

Marked as duplicate simplified: 38

(Codes listed under MARC8 is EACC)

MARC8	PUA	Character Name or Description	Alternate
213538	E900	East Asian ideograph (duplicate simplified)	台 (53F0)
273169	E901	East Asian ideograph (duplicate simplified)	系 (7CFB)
27322E	E902	East Asian ideograph (duplicate simplified)	仃 (4EC3)
273263	E903	East Asian ideograph (duplicate simplified)	尽 (5C3D)
273746	E904	East Asian ideograph (duplicate simplified)	当 (5F53)
273761	E905	East Asian ideograph (duplicate simplified)	罗 (7F57)
273D4F	E906	East Asian ideograph (duplicate simplified)	汇 (6C47)
274349	E907	East Asian ideograph (duplicate simplified)	历 (5386)
27457A	E908	East Asian ideograph (duplicate simplified)	欠 (6B20)
274E6F	E909	East Asian ideograph (duplicate simplified)	只 (53EA)
274F4B	E90A	East Asian ideograph (duplicate simplified)	获 (83B7)
274F70	E90B	East Asian ideograph (duplicate simplified)	巴 (5DF4)

275052	E90C	East Asian ideograph (duplicate simplified)	签 (7B7E)
275062	E90D	East Asian ideograph (duplicate simplified)	胡 (80E1)
275175	E90E	East Asian ideograph (duplicate simplified)	系 (7CFB)
275422	E90F	East Asian ideograph (duplicate simplified)	脏 (810F)
275458	E910	East Asian ideograph (duplicate simplified)	巴 (5DF4)
275551	E911	East Asian ideograph (duplicate simplified)	胡 (80E1)
275679	E912	East Asian ideograph (duplicate simplified)	胡 (80E1)
27574A	E913	East Asian ideograph (duplicate simplified)	冲 (51B2)
275E6B	E914	East Asian ideograph (duplicate simplified)	辟 (8F9F)
275F3E	E915	East Asian ideograph (duplicate simplified)	只 (53EA)
27605D	E916	East Asian ideograph (duplicate simplified)	台 (53F0)
27615F	E917	East Asian ideograph (duplicate simplified)	发 (53D1)
276163	E918	East Asian ideograph (duplicate simplified)	松 (677E)
276164	E919	East Asian ideograph (duplicate simplified)	胡 (80E1)
276165	E91A	East Asian ideograph (duplicate simplified)	须 (987B)
277258	E91B	East Asian ideograph (duplicate simplified)	恶 (6076)
283B7D	E91C	East Asian ideograph (duplicate simplified)	台 (53F0)
28702E	E91D	East Asian ideograph (duplicate simplified)	团 (56E2)
287271	E91E	East Asian ideograph (duplicate simplified)	纤 (7EA4)
287431	E91F	East Asian ideograph (duplicate simplified)	坛 (575B)
292433	E920	East Asian ideograph (duplicate simplified)	茫 (8298)
2D3C6D	E921	East Asian ideograph (duplicate simplified)	茫 (8298)
393B78	E922	East Asian ideograph (duplicate simplified)	峄 (5CC4)
4B5361	E923	East Asian ideograph (duplicate simplified)	角 (89D2)
4B5C54	E924	East Asian ideograph (duplicate simplified)	辟 (8F9F)
4B5E6C	E925	East Asian ideograph (duplicate simplified)	阝 (961D

Why Unicode?

Unicode is an international standard to encode all languages in the world correctly on the computer. It allows us to show several different languages on one web page providing that the correct support fonts (UTF-8) is installed and an appropriate browser (either Netscape 6.2+ or IE 6.0+) is used. Unicode is a very complex issue. There are still many more unresolved issues. What I have discovered from working with the Unicode™ Task

Force is just the tip of iceberg. This article intends to alert my cataloging colleagues to be aware that certain ideographs of EACC have mapped to the PUA code points, resulting in some missing characters when a record displays on OPAC. Hopefully, EISI will be using this new mapping file on its Unicode version of Voyager system. We will be able to see better display of our CJK records on OPAC. Of course, the best solution to this problem is to have the Unicode Consortium to submit more new Unicode proposals for new characters.